Stellenbosch Papers in Linguistics, Vol. 46, 2016, 77-97
doi: 10.5774/46-0-223

Omission and other sins: Tracking the quality of online machine translation output over four years

Susan Lotz

Language Centre, Stellenbosch University, South Africa
E-mail: slotz@sun.ac.za

Alta van Rensburg

Language Centre, Stellenbosch University, South Africa
E-mail: avrens@sun.ac.za

Abstract
Online machine translation (MT) has empowered ordinary language users to have texts translated all by themselves. But are these users aware of the pitfalls? This article draws on a longitudinal study that explored the quality of output by online MT application Google Translate in the language combination Afrikaans–English. We investigated the distribution of errors in two sets of translations (slide-show text and news report text) that we had Google Translate produce annually over a period of four years, 2010–2013. Omission, Mistranslation, Non-translation and Grammar were error categories that scored high in the analyses. In addition, we found that although the quality of the translations seemed to improve up to 2012, the pattern of improvement levelled off, with some of the 2013 output containing more errors than that of the previous year. We believe users should be made aware of the risks they unknowingly take when using online MT.

Keywords: error categories, Google Translate, machine translation, mistranslation, non-translation, translation quality

1. Introduction

Online machine translation (MT) has tipped the scale: translation is not reserved for translators anymore, but has become everyone's business. Garcia (2009: 205) illustrates the state of affairs when he says: "[Online] MT embodies the trinity of our brave new web world: free, instantaneous, and easy to use." The latest Google Translate mobile application is further testimony to this: any smartphone user can now use the phone's camera to have a sign in a foreign language translated at once – for some languages, without even having to take an actual photo. In January 2015, the official Google Translate blog estimated that more than 500 million people used Google Translate, in mobile app format or online, every month (Google 2015).

Google Translate is but one manifestation of how translation and technology connect constantly in new ways. Online MT makes it possible to find out within seconds what the gist of a text in

a foreign language is, potentially opening new worlds to its users (Doherty and O'Brien 2014: 40; Garcia 2009: 206; Hartley 2009: 121; Sager 1994: 262).

Not only the general public resorts to online MT – in a survey conducted in 2012, language professionals reported using free online MT (Gaspari, Almaghout and Doherty 2015: 14). The authors who conducted the 2012 survey also reported on the 20 top language combinations for which respondents used MT (Gaspari et al. 2015: 15; 17). The most frequent combination was English–French, followed by English–Italian and French–English, since the respondents originated mainly from Europe. The combination relevant to this study, Afrikaans–English, ranked 18th, since 50 of the 438 respondents were in fact South African.

Despite the increasing popularity of online MT among the general public and even translators, raw online MT output is rarely useful for more than gisting, due to the errors that occur in such translations. The output, as with other MT output, could be used as a starting point and post-edited to be improved to the level required, an approach whereby quality criteria are set by the purpose of the translation (Drugan 2013: 98; Garcia 2009: 206; Koponen and Salmi 2015: 119), but casual users of online MT do not necessarily know that.

Subsequently the question arises that, if one were to use the raw output of an online MT engine, what errors could be expected? Previous studies in this context have identified particularly two kinds of error that could compromise the transfer of meaning severely, namely omission and mistranslation.

The omission of certain words is a recognised high-risk practice in several MT systems, as DeCamp (2009) has pointed out with reference to Chang-Meadows' (2008) study of Chinese–English translations. DeCamp (2009) further remarks that uninformed users are not even always aware of what is omitted in the translations they obtain from online MT. In their research on the TC-STAR project with English–Spanish MT, Vilar, Xu, D'Haro and Ney (2006) have also identified "missing words" as a frequent error.

Mistranslation in MT has been flagged by Gaspari, Toral and Naskar (2011), who highlight the mistranslation of compounds in particular as a prominent error in German–English MT. Studies conducted in 2012 in language combinations such as Portuguese–English (Valotkaite and Asadullah 2012) and Spanish–English (Avramidis, Burchardt, Federman, Popović, Tscherwinka and Vilar 2012) also highlight mistranslation as the most frequent error encountered. In her error analysis of English–Finnish Google Translate output of three different text types, Koponen (2010: 7) found the most typical error to be "mistranslating an individual concept".

With a view to the above, we were interested in seeing how Afrikaans–English output of an online MT application would fare in an error analysis – also to be able to advise the general public better on this matter.

2. Context of this study

At the Stellenbosch University Language Centre[1], we observe an awareness of the new possibilities that the integration of language and technology opens to the students and staff of our university. Since the Centre renders a translation service, among others, questions regarding the use of free, online MT – and Google Translate, in particular – have been directed to us by clients ranging from students to staff and external clients. A frequent question posed is whether it would not be faster and more cost-effective than regular translation to employ online MT, given the need for the availability of texts in both Afrikaans and English in our higher education setting.

Translation is an everyday need at Stellenbosch University, as it is at many universities in South Africa. Despite the country's 11 national languages, English dominates nationally as the language of academic instruction and scholarship. Stellenbosch University functions as a multilingual institution that focuses on three languages: Afrikaans, English and isiXhosa (Stellenbosch University 2014). The University aims to accommodate students by offering learning opportunities in Afrikaans and in English, while developing and furthering isiXhosa[2]. Afrikaans is spoken by 13.5% of South Africans as their first language, while English is the first language of 9.6% of the population. IsiXhosa is spoken by 16% of South Africans as their first language. (South Africa.info 2012).

With a view to exploring the usability of online MT for University students who frequently rely on translation for both academic reading and writing, we began investigating the quality of translation products delivered by Google Translate in 2010 (see Lotz and Van Rensburg 2014; Van Rensburg, Snyman and Lotz 2012). The current article takes forward this ongoing research over a longer time span and with a further analysis of the distribution of errors. In this article, we report on (i) the distribution of errors in two sets of translations obtained from Google Translate annually over a period of four years, and (ii) whether the initial pattern of improvement in the quality of Google Translate output we saw in Lotz and Van Rensburg (2014) has continued since we added another kind of text and another year's output to the 2014 study results.

The next section will describe the methodology of our research, detailing the online MT application, texts, assessment tool and assessors involved. This is followed by a discussion of the results of our error analysis, focusing on the most frequent errors in each of the two texts, and commonalities in errors between the two kinds of text. In the concluding section, we summarise the findings of this study and discuss their implications for users of online MT.

[1] The purpose of the Stellenbosch University Language Centre is to render language support to students, staff and external clients who require such assistance. The Language Centre offers academic literacy modules, language acquisition modules, language planning expertise, document design services and language editing, translation and interpreting services. It also operates a writing laboratory and a reading laboratory.

[2] The current Stellenbosch University Language Policy applies until the end of 2016, and incorporates English in addition to Afrikaans as a language of instruction. A new language policy will take effect in January 2017, which will aim to ensure that no student will be excluded from the Stellenbosch University academic offering based on the student's command of Afrikaans or English. The 2017 policy explicitly makes provision for students who prefer to study in Afrikaans, while also improving access to education for students who are proficient in English only (Stellenbosch University 2016).

3. Methodology

At the end of October for four consecutive years, namely 2010, 2011, 2012 and 2013, we had two texts (text from a slide-show presentation and an online news report, see 3.2 for more detail) translated from Afrikaans to English by the online MT application Google Translate. At no stage did we use edit-and-improve suggestions that may have been offered by the application, and the source texts (STs) and resulting target texts (TTs) were not made public in any way. This means that the translated pairs could not re-enter Google Translate's database to be recycled. The process yielded eight translation products, which we analysed by means of an error analysis to determine the distribution of errors in the texts.

3.1 Translation application

Google Translate is the most widely-used free online translation application available currently (Drugan 2013: 170) and has been offering Afrikaans since September 2009 as one of the 103 languages into or from which it translated by July 2016 (Google n.d; 2009). Our clients enquired specifically about harnessing this application for translation between Afrikaans and English. Hence, Google Translate was used to produce the translations that were to be analysed.

Google Translate employs statistical machine translation (SMT) to compute the probability of what a translation would be (Kenny and Doherty 2014), after which it produces the translation with the highest probability. The computing process employs (i) translation models that have been trained on parallel corpora (an extraordinary large collection of STs and corresponding TTs translated by humans), and (ii) language models of the target language, which enables the system to check whether a certain combination of words is a likely sequence in the target language before it produces that sequence (Hearne and Way 2011).

Since the models reflect the data that were used to train them (Kenny and Doherty 2014: 284), it follows that, if one were to use Google Translate to translate a text in a subject field in which the system's language and translation models had no corpora to "learn" from, the results would be poorer than when the system translated something in which it had had training. The general public does not necessarily know or understand this. We concur with Kenny and Doherty (2014: 288) who observe that Google Translate actually may be "too easy to use" in that such systems "[obscure] the human labour that produces the translated and other data on which [SMT] is based; [they] also obscure the labour of the computer scientists who builds [SMT] systems". Free online translation gives the impression that translation is "an agentless, automatic function that can be realised in no time at all" (Cronin 2012: 47, in Kenny and Doherty 2014: 288), while that is not the case. We believe that users of free online MT may be deceived by how easy it is to obtain such translations, and that they may be unaware of the errors lurking in those translations.

3.2 Texts

We chose to work with two kinds of text that we encounter in our daily lives: slide-show text of a university lecture and an online news report. We regarded each of these texts as broadly representative of other texts similar to them in terms of function and form. With "kind of text", we thus mean what is called "*Textsorte*" in German, according to Kussmaul (1997: 69) and Snell-Hornby (1997: 278), and for the purposes of this study we distinguish between texts on

the basis of their specific formal and linguistic features and the particular situation in which they function. This makes it possible to distinguish between categories such as slide-show text, news reports, manuals, instruction leaflets, business letters, weather reports, cooking recipes, examination papers and minutes of a meeting.

In our experience, slide-show text and online news reports are two kinds of text that are often translated online in real-life situations – for example, over the past few years we had numerous students enquiring about using Google Translate to translate lecturers' slide-show texts, usually provided in either Afrikaans or English. Furthermore, Google Translate output of this particular slide-show text and news report respectively scored the highest and the lowest in an earlier phase of our study in which we evaluated six different kinds of text: a news report, minutes, an official letter, an examination paper and the slide-show text of a lecture. At the time, we had respondents evaluate the quality of the translations by means of a more holistic instrument than the method in the current study, namely an adaptation of Colina's (2008, 2009) evaluation instrument, as reported in Van Rensburg et al. (2012).

The slide-show ST was originally created as a Microsoft PowerPoint (MS PP) presentation in Afrikaans for a lecture in social anthropology and consisted of 312 words organised into 10 slides. The text in the presentation was copied from MS PP and pasted into an MS Word document, which was subsequently submitted to Google Translate. The text is characterised by short, bullet-like sentences that are dense with information. The correct translation of terminology would be key for this kind of text.

The online news report ST originated from an Afrikaans newspaper (in print and online) that circulates in the Western Cape Province, namely *Die Burger*, and consisted of 438 words. The online text was copied and pasted into an MS Word document, after which it was submitted to Google Translate. The text is characterised by full sentences, written in a typical journalistic style detailing the progress made in a case of disciplinary action against a politically active figure. Due to its political angle, the text contains many names. Therefore, it would be important that the names and dates are transferred correctly in the translation, and that the sentences in the translation are well formed. Rather than choosing a 312-word excerpt for the analysis so that it could be comparable with that of the 312-word slide-show text, we analysed the whole report. We wanted our analysis to be representative of typical online news reports, and analysing a 312-word excerpt would not serve the purpose. The fact that we worked with two texts with a word-count difference of more than 100 words had implications for the way in which the results could be compared. We took that into account.

3.3 Assessment tool

The evaluation of SMT output is a complex and contentious issue. Automatic metrics such as the Word Error Rate, the Position Independent Word Error Rate and the BLEU (Papineni, Roukos, Ward and Zhu 2001), NIST (Doddington 2002) and METEOR (Banerjee and Lavie 2005) evaluation metrics are prominent in SMT evaluation, since they give fast and cost-effective evaluations of (mostly evolving) translation models. However, their results are not failsafe. According to Callison-Burch (2009: 286), "they only loosely approximate human judgments", while Vilar et al. (2006) consider the interpretation of these measures as not clear at all. Daelemans and Hoste (2009: 9) hold that automatic evaluation measures "are only indirectly linked to translation usability and quality".

We chose not to work with automatic evaluation metrics, due to the mentioned concerns, and since the results would not have given us or our clients what we needed. Van Slype (1979, in Daems, Macken and Vandepitte 2013: 63) already argued more than 30 years ago that, since translation quality is not an absolute concept, it should be assessed "relatively, applying several distinct criteria illuminating each special aspect of the quality of the translation". Our analysis was a small one, and we needed an analysis by humans for humans, containing simple, practical examples of how the output of an online MT system looked, and what errors could be expected for texts similar to those we used in our study.

We therefore decided on a method that would usually be used to measure the quality of human translation – we adapted the Framework for Standardized Error Marking of the American Translators' Association (ATA) to perform the error analyses of the online MT output in our study. The ATA Framework is a "ready-made, standardised, time-tested, and professionally recognised model for conducting theory-based, systematic, coherent, and consistent" evaluation of translations (Doyle 2003: 21) and is used in the ATA certification examination (ATA 2015a,b). It enables the analyst to specify errors by type, which made it a useful evaluation tool for our purposes. We needed an instrument that could be applied as objectively as possible in the naturally subjective process of evaluating the quality of a translation product. Our error typology included categories such as Mistranslation, Addition, Omission, Non-translation, Switched elements, Terminology, Inconsistency, Grammar, Syntax, Word form, Spelling, Punctuation and Capitalisation.[3] Generally MT itself has no conception of such categories or other linguistic categories, as Kenny and Doherty (2014) argue, but our purpose in using the error analysis, as stated earlier, was to give prospective users of Google Translate insight into the possible errors they might encounter if they chose to translate texts similar to those we used in our study. Our framework for error marking is available as an appendix to this article.

3.4 Errors

With 'error' in this article, we mean that something in the translation output is wrong (Hansen 2010: 385). In translation studies, a distinction is often made between the kinds of error that occur in translated texts. Koby and Champe (2013: 165) distinguish between language errors, which entail "error[s] in the mechanics of target language usage", and translation errors, which concern errors in the "transfer of meaning". Correspondingly Pym (1992, 2010) differentiates between errors for which a clearly wrong and a clearly right option exist, calling them binary errors, and non-binary errors, for which there would be at least two right options in addition to the wrong option(s). Most language errors are binary errors, whereas there usually are multiple ways to correct translation errors, due to their non-binary nature.

In this study, all errors were marked and counted to establish how many errors there were and how frequently they occurred – regardless of whether those errors would be language errors or translation errors. In section 5, we use these scores to make several comparisons and observations regarding the distribution of errors in the two texts concerned. However, to make provision for how errors differ in gravity or severity regarding the influence they have on the meaning of a TT (Hansen 2010; Koby and Champe 2013: 165), we also assigned weights to the various errors to obtain an additional score: a weighted error score. Severe errors (that have a

[3] For a discussion of the error categories and how we adapted the Framework for our purposes, please see Lotz and Van Rensburg (2014).

significant impact on the transfer of meaning) were assigned a weight of 2, and less serious errors were weighted as 1. The weighted error score works counterintuitively: the higher the score, the lower the quality of the translation. We contrast the weighted error scores with scores obtained from the error analyses in the discussion of the results.

3.5 Assessors

The first author performed the error analysis on the two texts in question, and the second author verified them. At the time of the error analysis, the first author had a master's degree in general linguistics and 14 years' experience as a language practitioner. The second author had a master's degree in translation and was working on her PhD. She had 12 years' experience as a language practitioner. Both assessors also had experience in the revision, evaluation and assessment of translation products. Before performing the analyses, both assessors had done extensive reading on MT and had already investigated online MT by means of another evaluation instrument (Van Rensburg et al. 2012).

4. Results

The results of the assessment of the slide-show text will be described first, followed by those of the newspaper article. The categories we regarded as the most conspicuous, either due to significant changes in error counts over the four years or due to another reason, will be discussed for each text.

4.1 Errors in the slide-show text

The distribution of errors in the slide-show text translated by Google Translate over the four years in question is represented in Figure 1.

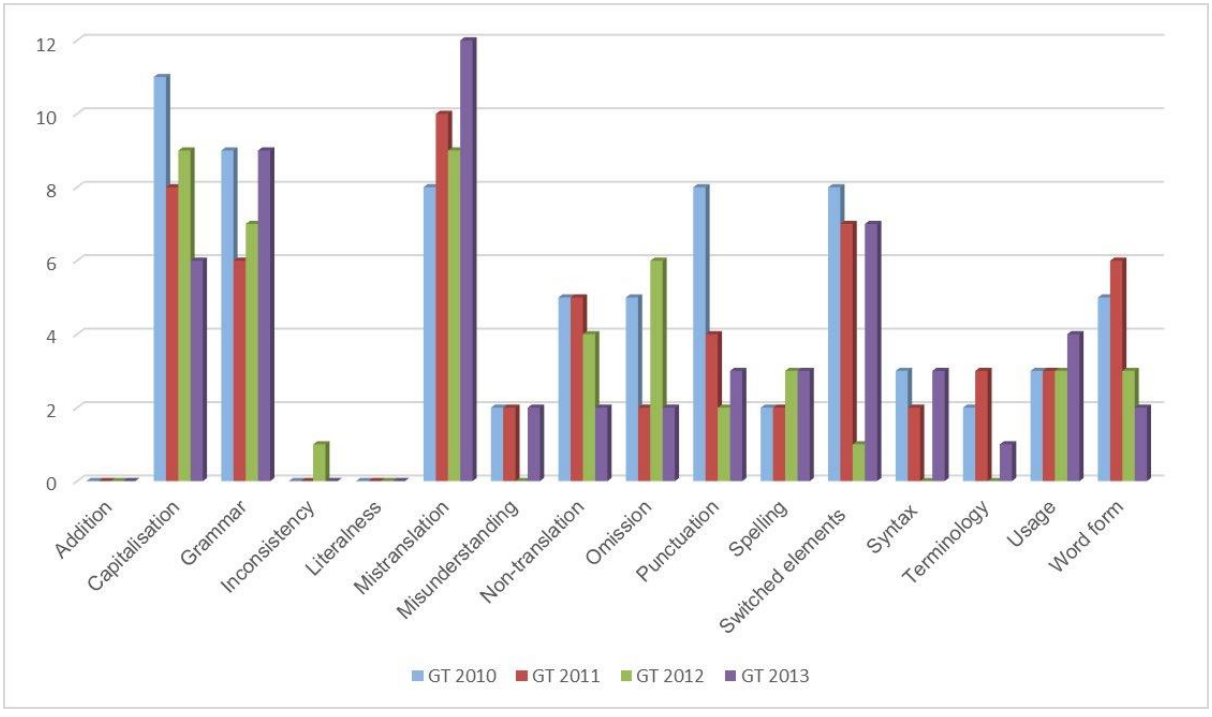

Figure 1: Distribution of errors in slide-show text output per year: 2010 to 2013

The highest number of errors in any category in any of the four years was recorded in the Mistranslation category for 2013. As mentioned in the introduction of the article, Koponen (2010: 7) found the most typical error to be "mistranslating an individual concept" in her error analysis of English–Finnish Google Translate output of three different text types. In our analysis, Mistranslation errors increased from 8 in 2010 to 10 in 2011, decreased by 1 in 2012 and shot up by 3 to make a total of 12 mistranslations in 2013. Mistranslation in the 2013 text constitutes 21,5% of all errors in that text – a large percentage, particularly in light of the serious impact that mistranslation could have on the quality and trustworthiness of a translated text. As already mentioned, in an earlier part of our study (Van Rensburg et al. 2012) we found that, of the six kinds of text that we had previously investigated, Google Translate had performed best when translating slide-show text. Thus, the application still did not perform well in this crucial category even when translating a kind of text that we found more suitable than other kinds of text for online MT. It could be argued that the system was not trained to perform optimally when translating this kind of text or in this subject field. However, that was exactly the point of this study: to simulate the circumstances under which a member of the general public would run a text through a free online MT system – a system that was not trained specifically for the text it was used for.

As reported in Lotz and Van Rensburg (2014), a major mistranslation in the 2010 to 2012 output occurred in the following instance:

> Afrikaans ST: *bier, handel en buite-egtelike verhoudings*.
> Benchmark English TT: beer, trade and extramarital affairs.

The 2010 to 2012 Google Translate output for the above phrase is presented in Table 1.

Table 1: 2010 to 2012 Google Translate output: "buite-egtelike verhoudings" (Lotz and Van Rensburg 2014)

Year of output	**Output**
2010	beer, trading and extra-marital ***relations**
2011	beer, *marketing and extra-marital ***relationships**
2012	beer, trade and ***foreign** affairs

*denotes an error

We have since analysed the 2013 output, which showed that Google Translate recovered from its 2012 "blunder" (seen from a user perspective) in its 2013 output:

> 2013: beer, trade and extra-marital *relationships

However, the string "extramarital relationships", which was also produced in 2011, still is not quite an adequate translation for what should have been "extramarital affairs". Therefore, we marked it as a Mistranslation error in the error analysis.

Now consider an example of Non-translation that evolved into a Mistranslation error. Table 2 contains the relevant Google Translate output in the four years under review.

> Afrikaans ST: *Rituele mistifiseer die rol van vroue*.
> Benchmark English TT: Rituals mystify the role of women/Ritual mystifies the role of women.

Table 2: 2010 to 2013 Google Translate output: "mistifiseer"

Year of output	Output	Error marked
2010	Ritual ***mistifiseer** the role of women	Non-translation
2011	Ritual ***mistifiseer** the role of women	Non-translation
2012	Ritual ***demystify** the role of women	Mistranslation, concord
2013	Rituals ***demystify** the role of women	Mistranslation

Google Translate did not translate the word "mistifiseer" in 2010 and 2011. The untranslated word was copied into the translated text, without further processing. Failure to translate this word was marked as a Non-translation error. In 2012 and 2013, the application used "demystify" as a translation equivalent for "mistifiseer" in the ST, which means exactly the opposite. Consequently, it was marked as a Mistranslation error.

The error count for Capitalisation was also quite high in 2010 but has decreased sharply over the four years. Since capitalisation does not influence the transfer of meaning in an Afrikaans–English setup significantly, we will not discuss that category further here.

Grammatical errors, starting at nine errors, initially decreased by three from 2010 to 2011, but then increased to seven errors in 2012 and to nine in 2013, thereby equalling the initial error count. Despite the fluctuation in the number of errors there seems to have been no overall improvement on the grammar front over the four years.

An important category concerning the transfer of meaning that reflected improvement over the four years is Non-translation. This is a black-or-white category in the sense that a word is either translated or not – there are no nuances that may have influenced the identification of this kind of error. We added this category to our framework to provide for a common Google Translate error mentioned earlier, namely that, if the application does not find a match for a ST word or combination of words, it simply copies the ST word into the TT. In 2010, five non-translations were recorded, with no improvement in 2011, but 2012 yielded four errors and 2013 only two.

An example of non-translation is Google Translate's dealing with "deelsaaiery", of which the benchmark translation would be "share-cropping". In 2010 and 2011, the application did not translate this word, inserting the untranslated word in the TT. In 2012, it used "share – cropping", but used spaces and an en dash instead of a hyphen, which means that typographically the translation was still incorrect. In 2013, it reverted back to not translating the word and copying "deelsaaiery" in the TT.

The last category to be highlighted in Figure 1 is Switched elements – another category that we added to our framework to provide for a frequent error in Google Translate output, in our experience. "Elements" may refer to words or phrases. This category involves two adjacent elements that were translated correctly in the TT, but that appear to be switched around, in comparison to their position in the ST (Lotz and Van Rensburg 2014). An example of an error in this category is presented in Table 3. (Although there are many errors in the output in this example, we focus only on Switched elements here).

> Afrikaans ST: *Periferie = buite die kern, uitgebuite gebied*
> Benchmark English TT: Periphery = outside the core, exploited area

Table 3: 2010 to 2013 Google Translate slide-show text output: Switched element error

Year of output	Output	Error marked
2010	= Periphery outside the core area exploited	Switched elements
2011	= outer periphery of the core area exploited	Switched elements
2012	Periphery = exploited area outside the nucleus	–
2013	= Periphery outside the core area exploited	Switched elements

In the 2010 and 2013 output in Table 3, the first two elements, namely "periphery" and "=" have been switched, which influences the transfer of meaning significantly. In the ST and the 2012 output, the text that follows the equation mark serves as a definition of what precedes the equation mark, whereas the 2010, 2011 and 2013 output incorporate what preceded the equation mark (in the ST) in the nonsensical definition that follows the equation mark. In this example, an error of Switched elements results in no term to be defined as well as a nonsensical definition in the TT.

Switched elements in the slide-show text started off with eight errors in 2010, decreased by one in 2011 and then dropped to only one error in 2012. Then, in 2013, the count rose to seven again, constituting 12.7% – the third highest score – of all errors in 2013. The increase in Switched element and Mistranslation errors in 2013 influenced the year's total error score greatly, as will be shown later.

4.2 Errors in the news report

The distribution of errors in the translated news report text over the four years in question is represented in Figure 2.

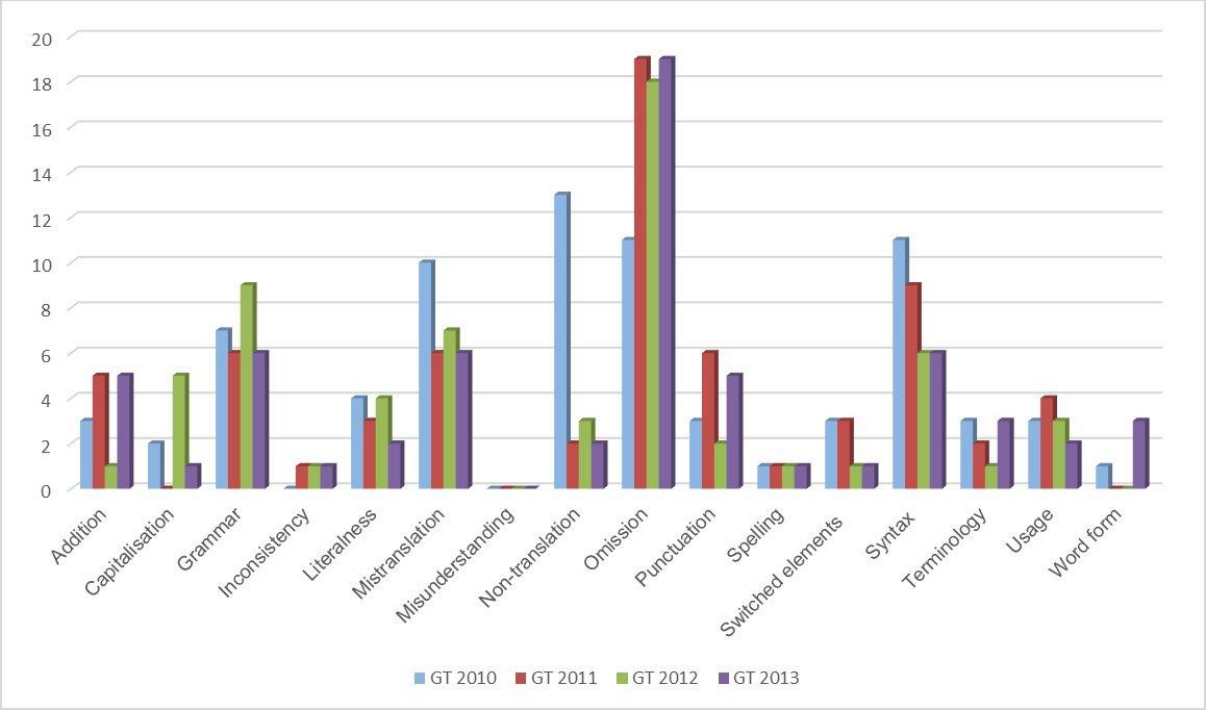

Figure 2: Distribution of errors in online news report output per year: 2010 to 2013

The most conspicuous error category in Figure 2 is Omission. In 2010, it started off with a score of 11 errors and then increased sharply to 19 errors in 2011. The error count seems to have stabilised with 18 errors in 2012 and once again 19 errors in 2013. If one regards Omission in isolation, there is no sign of improved quality in this text, but there are other categories to consider.

The pattern of improvement that we saw in Non-translation in the slide-show text analysis is repeated in the news report analysis. In 2010, 13 Non-translation errors were recorded, which decreased dramatically over the four years to only 2 in 2013. This improvement is in accordance with the claim by the developers of Google Translate that the application would improve over time (Helft 2010). From our results, it would seem that the Google Translate vocabulary has definitely improved. However, one should, keep in mind that Non-translation errors often evolve into other errors, as shown in the discussion of Non-translation errors in the analysis of the slide-show text.

Overall, Mistranslation errors also decreased markedly, although the improvement from 2010 (10 errors) to 2011 (6 errors) levelled off with 6 errors in 2013. Tables 4, 5 and 6 set out a few examples of Mistranslation. The applicable ST and benchmark TT precede the table in each instance.

Table 4 illustrates Google Translate's inefficiency when translating a title that seems to not have occurred in its training data. "Me." in Afrikaans means "Ms" in English, but Google Translate translated "Me." with "sorry." in 2010 and 2011. In 2012 and 2013, the title was simply not translated and the Afrikaans "Me." was reproduced in the translation. We marked this as Non-translation.

> Afrikaans ST: *Me. Ayanda Dlolo*
> Benchmark English TT: Ms Ayanda Dlolo

Table 4: Example of Mistranslation error in news report translation output: title

Year of output	**Output**	**Error marked**
2010	*sorry. Ayanda Dlolo	Mistranslation
2011	*sorry. Ayanda Dlolo	Mistranslation
2012	*Me. Ayanda Dlolo	Non-translation
2013	*Me. Ayanda Dlolo	Non-translation

A mistranslated surname came up where "Mr Mathews Phosa" in the ST became "Mr Slabbert" in the 2012 TT, as shown in Table 5. Seeing that Mr Phosa is a well-known politician and that the surname Slabbert is also associated with a well-known politician in South African politics, this mistranslation is somewhat ironic in a South African context. In the following year, the surname was again as it should be, namely "Phosa".

> Afrikaans ST: *[Mnr Mathews] Phosa*
> Benchmark English TT: [Mr Mathews] Phosa

Table 5: Example of Mistranslation error in news report translation output: surname

Year of output	Output	Error marked
2010	Phosa	–
2011	Phosa	–
2012	*Slabbert	Mistranslation
2013	Phosa	–

We also came across the mistranslation of an indication of time, as shown in Table 6. "Vandag" in the Afrikaans ST, which should have been translated as "today", was translated as "on Monday" in 2013. What we found interesting here is that the specific day in 2010 on which that news report was published was in fact a Monday. Table 6 contains the output in the different years.

> Afrikaans ST: *Volgens die ANC moet Malema hom vandag by Luthuli-huis aanmeld…*
> Benchmark English TT: According to the ANC, Malema should report at Luthuli House today…

Table 6: Example of Mistranslation error in news report translation output: indication of time

Year of output	Output	Error marked
2010	Today	–
2011	Today	–
2012	Today	–
2013	on Monday	Mistranslation

Lastly, the name of the newspaper from which this report originated, *Die Burger*, was mistranslated in each of the four years under review, as shown in Table 7.

> Afrikaans ST: *City Press*, *susterkoerant van Die Burger*…
> Benchmark English TT: *City Press*, sister paper of *Die Burger*…

Table 7: Example of Mistranslation error in news report translation output: name of newspaper

Year of output	Output	Error marked
2010	The Citizen	Mistranslation
2011	The Citizen	Mistranslation
2012	The Argus	Mistranslation
2013	The Citizen	Mistranslation

"Citizen" may have been a good translation equivalent for "burger" if one wished to translate the word in its regular sense, but a proper name should not be translated. Moreover, a South African newspaper going by the name *The Citizen* in fact exists. In 2012, Google Translate used "The Argus" as a translation, which is in part the name of another existing South African newspaper, *Cape Argus*, and that of a UK newspaper. None of the papers *The Citizen*, *Cape Argus* or *The Argus* are translated versions of *Die Burger*. Errors of this nature underscore the fact that the system performing this translation did not have the contextual reference and agency to come up with a solution for this translation challenge.

Errors with regard to Literalness, a category closely related to Mistranslation in that both have a direct bearing on the transfer of meaning, follow more or less the same pattern as Mistranslation errors, albeit on a smaller scale: starting at four errors in 2010, decreasing to three in 2011, increasing by one again in 2012 to decrease to only two errors in 2013.

The Syntax category showed significant improvement over the four years, with an initial error count of 11 in 2010, decreasing to 9 in 2011, decreasing further to 6 in 2012 and levelling off with 6 errors in 2013. Syntax may well be regarded as more of a language error than a translation error, but syntax has a marked influence on the readability – and therefore usability – of a translated text. Percentage-wise there were more Syntax errors in the news report than in the slide-show text to start off with in 2010. This could be expected, as a news report consists of running text, while a slide-show contains factual information with a simpler structure. Nevertheless, the improvement regarding Syntax errors in running text specifically is noteworthy. Consider the example of an error-free, acceptably translated sentence that Google Translate produced in Table 8:

Afrikaans ST: *Dié optrede sal geskied onder leiding van mnr. Derek Hanekom, adjunkminister van wetenskap en tegnologie en die voorsitter van die tugkomitee.*

Table 8: Example of acceptable sentence in Google Translate output in all four years

Year of output	Output
2010	This action will take place under the **guidance** of Mr Derek Hanekom, deputy minister of science and technology and the chairman of the disciplinary committee.
2011	This action will take place under the **direction** of Mr Derek Hanekom, deputy minister of science and technology and the chairman of the disciplinary committee.
2012	This action will take place under the **leadership** of Mr Derek Hanekom, deputy minister of science and technology and the chairman of the disciplinary committee.
2013	This action will take place under the **guidance** of Mr Derek Hanekom, deputy minister of science and technology and the chairman of the disciplinary committee.

The English TT produced by Google Translate over the four years differ only regarding the translation of one word, namely "leiding". All the translation equivalents offered – "guidance", "direction" and "leadership" – were acceptable in this context.

5. Discussion

In the slide-show text analysis Mistranslation, Capitalisation, Grammar, Non-Translation and Switched elements were the categories that represented the most prominent errors and in which there was significant variation in the number of errors over the four years. In the news report analysis Omission, Non-Translation, Mistranslation, Literalness and Syntax accounted for the most prominent errors and error movement in that text. Non-translation and Mistranslation are categories that scored high in the analyses of both texts – two categories that have a major impact on the transfer of meaning.

5.1 Error category totals

An additional way to gain insight into the distribution of errors is to consider pie charts representing the results of the error analyses. Figure 3 shows the number of errors counted for each category in the slide-show text over the four years.

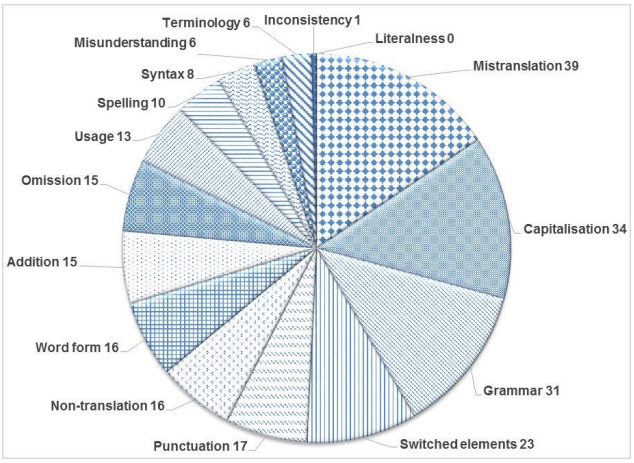

Figure 3: Number of errors in each error category for slide-show text output: 2010 to 2013

The four most prominent error categories over the four years under review in the slide-show text were: Mistranslation (39 errors), followed by Capitalisation (34 errors), Grammar (31 errors) and Switched elements (23 errors).

Contrast these numbers to those in Figure 4, which shows the distribution of errors in specifically the 2013 slide-show output.

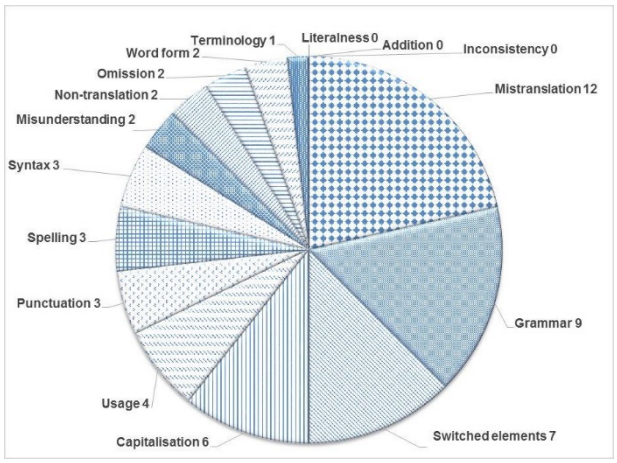

Figure 4: Number of errors in each error category for slide-show text output: 2013

Since the 2013 output is the most recent output in the study, Figure 4 gives the best overview of the kinds of error users of Google Translate could expect when having a similar text translated. In the 2013 slide-show output Mistranslation, Grammar, Switched elements and Capitalisation were still the categories with the highest error scores, and therefore the four most likely areas in which errors could be expected in future translations. Compared to the 2010 to 2013 totals, Mistranslation stays the greatest concern for this text.

Figure 5 shows the number of errors recorded for each category in the news report text over the four years.

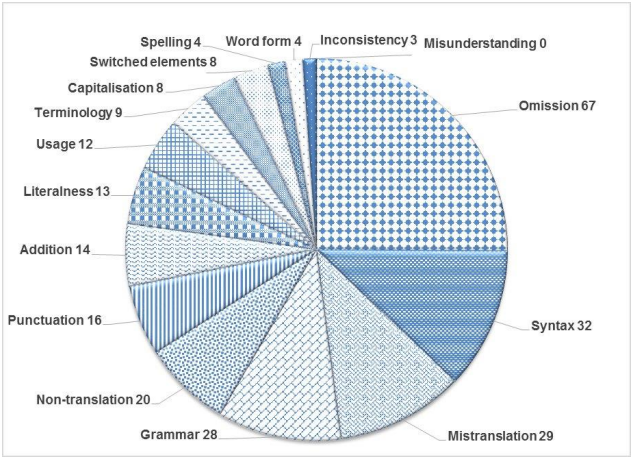

Figure 5: Number of errors in each error category for news report output: 2010 to 2013

In the news report output over the four years, the category of Omission had the most errors: 67. Syntax was second (32), followed by Mistranslation (29) and Grammar (28). In Figure 6, the isolated 2013 output results show that the four most likely areas in which errors would occur if Google Translate were to be used to translate a news report are the same: Omission, Syntax, Mistranslation and Grammar.

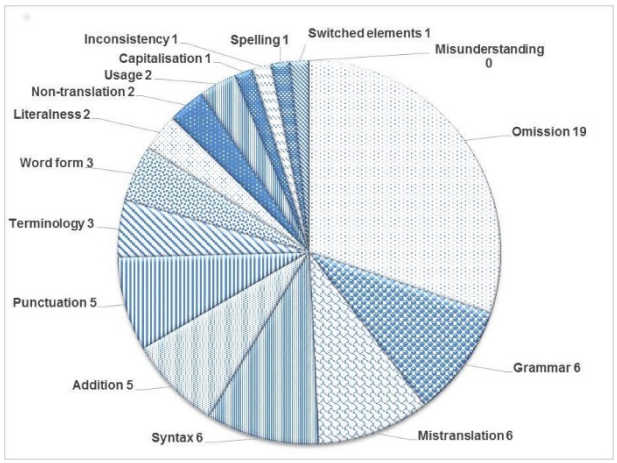

Figure 6: Number of errors in each error category for news report output: 2013

As explained before, the error category totals of the two texts should not be compared at face value, since the news report consisted of more words than the slide-show text. A longer text carries the possibility of more errors, therefore the aim of figures 3 and 5 is to give an overview of the errors that were encountered in each text type over the four years of investigation. What becomes clear from juxtaposing those two figures is that the error categories Mistranslation and Grammar were prominent in both analyses.

5.2 Distribution of errors over the four years

In figures 1 and 2, which show the number of errors in each category per year for each text by means of bar charts, the 2012 translations of both texts registered more errors in several categories than the 2011 translations. Also, despite improvement in some categories in the 2013 translations, there was also an increase in errors in quite a few categories. In the analyses, we observed that new errors were made in later years. Some errors made in earlier years were resolved, but then new ones would appear in the following year's translation – often elements that had in fact been correct in previous translations.

This observation brings a new question to mind: How does the distribution of total number of errors over the four years for each text compare to each other? To compare the error totals of the two analyses and the results for the different years, the total error score for each year of each text had to be converted to a percentage. We obtained percentage values by dividing the total error score of each year's analysed text by the word count of the applicable ST. The results of this process are shown in Figure 7.

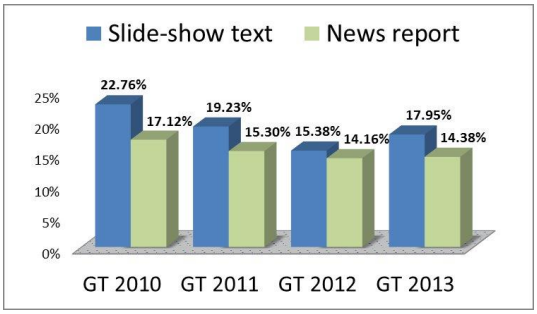

Figure 7: Totals of both text types as percentages for comparison

Figure 7 reflects a steady decrease in total errors in both texts up to 2012. However, both 2013 texts contained more errors than the texts produced in the previous year. The increase in errors in the 2013 slide-show text was much more pronounced, containing eight more errors, than in the news report, which contained only one more error than the previous year's text.

The weighted error scores (as introduced in the methodology section) also reflect an increase in 2013, with a more pronounced increase in the 2013 slide-show text than in the news report. Consider Table 9 in this regard, keeping in mind that the higher the score, the more questionable the quality of that translation.

Table 9: Error scores versus weighted error scores

		2010	2011	2012	2013
Slide-show text	Number of errors	71	60	48	56
	Weighted error score	107	92	69	86
News report	Number of errors	75	67	62	63
	Weighted error score	111	95	88	87

We believe that the new errors in the latest translations in particular, may be a reflection of the unpredictable quality of the corpora from which Google Translate draws its translation equivalents. New bilingual corpora could be sourced from virtually anywhere, influencing the

quality of the translation equivalents the system offers positively or negatively. Therefore, it may be that Google Translate sometimes seems to revert to mistakes made in earlier years, even if it used an acceptable translation equivalent for a while, because new data (which could also be recycled data) are still being added to the system. The stream of new data is declining; however, and Macduff Hughes, engineering director of Google Translate, acknowledged in an interview published in April 2015 that, "for the most common language pairings, 'we have reached about the limit where more data is helpful'" (Greene 2015: 33). The system needs a radical breakthrough for a new leap forward in quality. Until this breakthrough, we suspect the quality of Google Translate output will not necessarily continue improving over time as it did initially – by 2013 it had levelled off for the texts we analysed, for example.

This is a very small study in, globally seen, a very small language pair. However, there are similarities with what researchers found in other studies on MT. Omission and mistranslation, the categories mentioned in the introduction of this article, are also prominent in our study in that mistranslation stood out as the greatest concern in particularly the slide-show text, and 30% of all errors identified in the 2013 news report text were errors of omission. In light of these findings, casual users of online MT need to be made aware of the probability of particularly mistranslation and omission in the translations they obtain from online MT systems.

6. Conclusion

The results of the current study confirm the present dialogue on the quality of online MT, summarised in Gambier's (2014: 11) statement that "the translations produced by Google Translate, for example, are of good enough quality because they are consulted rather than actually read or assimilated". Online MT that has not been trained for translation in a specific subject field could be useful – but within certain parameters. We are concerned that casual users of online MT, particularly students and staff in our case, do not necessarily use online MT within those parameters, since they are not sufficiently aware of the high probability of errors in online MT translations. Casual users often lose sight of the fact that the system performing the translation lacks agency and does not automatically have the contextual reference they may take for granted.

This article presented the results of a study conducted to investigate the distribution of errors in two sets of translations (slide-show text and news report text) produced by Google Translate annually over a period of four years, 2010 to 2013. What we found was that the error categories Non-translation and Mistranslation – which have a major impact on the transfer of meaning – and Grammar scored high in the analyses of both texts. Other studies confirmed mistranslation (the highest scoring category for the slide-show text) and omission (the highest scoring category for the news report text) as high risks in MT.

In addition, we wanted to determine whether the initial pattern of improvement in the quality of Google Translate output that we had seen in Lotz and Van Rensburg (2014) would continue after we had added another kind of text and another year of output for the purposes of the current study. We found definite improvement in quality in the Google Translate output of the first three years under investigation. However, there were more errors in the output of the last year (2013) in both texts than in the output of the previous year (2012). From the error analyses, it seemed that new errors had been introduced in the 2013 translations. The improvement that was observed over the first three years of the study thus levelled off in 2013. The results of this

study confirm the observation of Lotz and Van Rensburg (2014: 248–9) that the very quality that enables Google Translate to improve dramatically over a span of time, its data-drivenness, also seems to make the application unpredictable and might hamper its progress.

Subsequent analyses of Google Translate output of 2014, 2015 and years to come of the same and possibly additional kinds of text may shed more light on whether the application's prowess improves further. Studies on the post-editing effort required for similar texts created by online MT in the language combination Afrikaans–English (and vice versa) in particular, would also be meaningful.

What the findings of this study mean for translators and the translate-it-yourself public alike is that it confirms that using an online translation application like Google Translate is a risk. Users of free online MT may not always be aware of (or qualified to determine) what has, for example, been omitted or mistranslated in the resulting TT. They may be deceived by how easy it is to obtain translations, and be unaware of the errors lurking in those translations. When they choose to use online MT, they should be educated enough to take a calculated risk. In this regard, Vitek (2000), an American freelance technical translator, had the following insight in 2000 already: "[i]t is up to us, translators, to explain to the general public what machine translation is, what are its strengths and weaknesses, and what is its likely role in the future development of our civilization".

References

American Translators Association (ATA). 2015a. *Certification*. Available online: http://www.atanet.org/certification/landing_about_exam.php (Accessed 14 April 2015).
American Translators Association (ATA). 2015b. *Framework for Standardised Error Marking: Explanation of error categories.* Available online: http://www.atanet.org/certification/aboutexams__error.php (Accessed 14 April 2015).

Avramidis, E., A. Burchardt, C. Federman, M. Popović, C. Tscherwinka and D. Vilar. 2012. Involving language professionals in the evaluation of machine translation. In *Proceedings of the Language Resources and Evaluation Conference*, 21–27 May, Istanbul, Turkey. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/294_Paper.pdf (Accessed 15 April 2013).

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, USA. pp. 65–73. Available online: http://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf (Accessed 21 June 2015).

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, 6–7 August. pp. 286–295. Available online: http://www.aclweb.org/anthology/D/D09/D09-1030.pdf (Accessed 10 April 2014).

Chang-Meadows, S. 2008. MT errors in CH-to-EN MT systems: User feedback. In *MT at work, Proceedings of the Eighth Conference of the Association for Machine Translation in the*

Americas, 21–25 October, Waikiki, Hawaii. Available online: http://www.mt-archive.info/ AMTA-2008-Chang-Meadows.pdf (Accessed 28 October 2013).

Colina, S. 2008. Translation quality evaluation: Empirical evidence for a functionalist approach. *The Translator* 14(1): 97–134.

Colina, S. 2009. Further evidence for a functionalist approach to translation quality evaluation. *Target* 21(2): 235–264.

Daelemans, W. and V. Hoste (Eds). 2009. Evaluation of Translation Technology. *Special issue of Linguistica Antverpiensia New Series – Themes in Translation Studies*, Vol 8. Antwerp: ASP.

Daems, J., L. Macken and S. Vandepitte. 2013. Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, September 2, eds. S. O'Brien, M. Simard and L. Specia. pp. 63–71.

DeCamp, J. 2009. What is missing in user-centric MT? In *Proceedings of MT Summit XII*, 26–30 August, Ottawa, Ontario, Canada. Available online: http://www.mt-archive.info/MTS-2009-DeCamp-1.pdf (Accessed 22 July 2014).

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, CA, USA. pp. 138–145. Available online: http://dl.acm.org/citation.cfm?id=1289273 (Accessed 15 August 2015).

Doherty, S. and S. O'Brien. 2014. Assessing the usability of raw machine translated output: a user-centered study using eye tracking. *Intl. Journal of Human–Computer Interaction* 30: 40–51. doi: 10.1080/10447318.2013.802199.

Doyle, M. (2003). Translation pedagogy and assessment: Adopting ATA's Framework for Standard Error Marking. The *ATA Chronicle*, November/December. pp. 21–29.

Drugan, J. 2013. *Quality in professional translation*. *Assessment and improvement*. London: Bloomsbury.

Gambier, Y. 2014. Changing landscape in translation. *International Journal of Society, Culture and Language* 2(2): 1–12.

Garcia, I. 2009. Beyond translation memory: Computers and the professional translator. *The Journal of Specialised Translation* 12, July, pp. 199–213. Available online: http://www.jostrans.org/ issue12/art_garcia.php (Accessed 27 February 2012).

Gaspari, F., H. Almaghout and S. Doherty. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives: Studies in Translatology*. Advance online publication. doi: 10.1080/0907676X.2014.979842.

Gaspari, F., A. Toral and S. Naskar. 2011. User-focused task-oriented MT evaluation for wikis: A case study. In *Proceedings of the Third Joint EM+/CNGL Workshop 'Bringing MT to the User: Research Meets Translators'*, 14 October 2011. Luxembourg: European Commission.

Google. N.d. Your world. Now with 103 languages. Available online: http://translate.google.co.za/about/intl/en_ALL/languages.html (Accessed 7 July 2016).

Google. 2009. Google Translate: The Official Blog. Google Translate now available for Swahili and Afrikaans, 3 September. Available online: http://google-africa.blogspot.com/2009/09/google-translate-now-available-for.html (Accessed 8 September 2009).

Google. 2015. Google Translate: *The Official Blog. Hallo, hola, ola more powerful Translate app*, 14 January. Available online: http://googleblog.blogspot.com/2015/01/hallo-hola-ola-more-powerful-translate.html (Accessed 4 February 2015).

Greene, L. 2015. Everything you ever wanted to know about Google Translate, and finally got the chance to ask. *TAUS Review of language business and technology*, 3: April, pp. 23–33. Available online: http://issuu.com/tausreview/docs/tausreview-dataissue-april2015/32 (Accessed 5 May 2015).

Hansen, G. 2010. Translation errors. In Y. Gambier and L. van Doorslaer (Eds). *Handbook of Translation Studies*. Available online: http://benjamins.com/online/hts/ (Accessed 6 February 2012).

Hartley, T. 2009. Technology and translation. In J. Munday (Ed). *The Routledge companion to translation studies*. New York: Routledge. pp. 106–127.

Hearne, M. and A. Way. 2011. Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass* 5(5): 205–226. doi: 10.1111/j.1749-818x.2011.00274.x.

Helft, M. 2010. Google's computing power refines translation tool. *The New York Times*, 9 March. Available online: http://www.nytimes.com/2010/03/09/technology/09translate.html (Accessed 4 February 2013).

Kenny, D. and S. Doherty. 2014. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer* 8(2): 276–294. doi: 10.1080/1750399X.2014.936112.

Koby, G. and G. Champe. 2013. Welcome to the real world: Professional-level translator certification. *Translation and Interpreting* 5(1): 156–172.

Koponen, M. and L. Salmi. 2015. On the correctness of machine translation: A machine translation post-editing task. *Journal of Specialised Translation* 23: 118–136. Available online: http://www.jostrans.org/issue23/art_koponen.php (Accessed 5 February 2015)

Koponen, M. 2010. Assessing machine translation quality with error analysis. In *Electronic proceedings of the KäTu symposium on translation and interpreting studies* 4. Available online: http://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf (Accessed 7 November 2014).

Kussmaul, P. 1997. Text-type conventions and translating. In A. Trosborg (Ed). *Text typology and translation*. Benjamins Translation Library. Amsterdam: John Benjamins. pp. 24–41.

Lotz, S. and A. van Rensburg. 2014. Translation technology explored: Has a three-year maturation period done Google Translate any good? *Stellenbosch Papers in Linguistics PLUS* 43: 235–259.

Papineni, K., S. Roukos, T. Ward and W-J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Technical report RC22176. Yorktown Heights, NY: IBM Research Division, Thomas J. Watson Research Center. Available online: http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf (Accessed 11 August 2013).

Pym, A. 1992. Translation error analysis and the interface with language teaching. Available online: http://usuaris.tinet.cat/apym/on-line/training/1992_error.pdf (Accessed 7 August 2013).

Pym, A. 2010. Text and risk in translation. Available online: http://usuaris.tinet.cat/apym/online/translation/risk_analysis.pdf (Accessed 12 October 2013).

Sager, J. 1994. *Language engineering and translation: Consequences of automation*. Amsterdam: John Benjamins.

Snell-Hornby, M. 1997. Written to be spoken. In A. Trosborg (Ed). *Text typology and translation*. Benjamins Translation Library. Amsterdam: John Benjamins. pp. 277–90.

South Africa.info. 2012. South Africa's population. Available online: http://www.southafrica.info/about/people/population.htm#.U8qC__mSy8A (Accessed 12 June 2015).

Stellenbosch University. 2014. *Language Policy of Stellenbosch University*. Available online: http://www.sun.ac.za/english/Documents/Language/Language%20Policy%202014%20Final%2012%20Dec%202014.pdf (Accessed 7 July 2016).

Stellenbosch University. 2016. *Language Policy of Stellenbosch University*. Available online: http://www.sun.ac.za/english/Documents/Language/Final%20Language%20Policy%20June%202016.pdf (Accessed 7 July 2016).

Valotkaite, J. and M. Asadullah. 2012. Error detection for post-editing rule-based machine translation. In *Proceedings of the Association for Machine Translation in the Americas 2012, Workshop on Post-editing Technology and Practice* (WPTP 2012, 28 Oct–1 Nov, San Diego, USA. Available online: http://amta2012.amtaweb.org/AMTA2012files/html6/6_paper.pdf (Accessed 13 April 2013).

Van Rensburg, A., C. Snyman and S. Lotz. 2012. Applying Google Translate in a higher education environment: Translation products assessed. *Southern African Linguistics and Applied Language Studies* 30(4): 511–524.

Vilar, D., J. Xu, L. D'Haro and H. Ney. 2006. Error analysis of machine translation output. In *Proceedings of the 5th LREC*, Genoa, Italy. pp. 697–702.

Vitek, S. 2000. Reflections of a human translator on machine translation. *Translation Journal* 4(3): July. Available online: http://www.bokorlang.com/journal/13mt.htm (Accessed 31 January 2012).